Review of the Stage 1 RR "Does truth pay? Investigating the effectiveness of the Bayesian Truth Serum with an interim payment"

The Stage 1 report under review proposes an experimental investigation of the effectiveness of an extended version of the Bayesian Truth Serum (BTS) for eliciting (more) truthful responding to sensitive questions. Specifically, the authors argue that mixed evidence regarding the BTS's validity may be due to a lack of trust in the procedure. They propose an online experiment with a 3-level between-participants factor Condition (Regular Incentive, RI; BTS; BTS + Interim Payment, BTS + IP). In each condition, participants respond to a set of 10 items asking about sensitive topics across 2 sessions (5 items per session). Based on the more-is-better assumption, the authors formulate three hypotheses implying the following pattern of average endorsement (i.e., mean ratings across items) across conditions:

$$\mu_{RI} < \mu_{BTS} < \mu_{BTS+IP}.$$

The challenge of eliciting truthful responses in self-reports on sensitive topics is an important one. In my own research, I work with randomized response models, and I am quite sympathetic to the authors' notion that psychological aspects of the questioning procedure – such as understanding and trusting the instructions – play a central role. I agree that this is particularly important in the context of the BTS that is based on quite sophisticated assumptions about how rationally participants behave in the survey situation. Including an interim payment is an interesting idea, and a thoroughly planned, registered experimental comparison with a regular BTS and a non-BTS condition is a welcome addition to the literature, in my view. Thus, my overall evaluation of this Stage 1 report is positive. Nevertheless, I want to point out some aspects that may need clarification, as well as concrete recommendations for (potential) improvement.

Major issues:

1) Justification of hypotheses

I agree with the authors that (a lack of) trust in the BTS's procedure might be an issue. Based on this assumption, the hypothesis that an interim payment may increase trust – and, in turn, truthful responding – is plausible. I wonder, however, whether interim payments could also have a negative effect on some participants. Assume that a participant provides, at least in their view, truthful responses. Yet, based on the i-score ranking, this person is not among the top group and receives no bonus payment. They may conclude that the algorithm to detect truthful responding isn't working properly or that the feedback is corrupt, potentially resulting in decreased trust in the procedure. The authors clearly don't expect this effect and they may have good reason for this – however, I recommend that they spell out these reasons more explicitly and justify their expectations/hypotheses more carefully.

2) Measures

2.1) The authors mention the i-score as their measure for ranking and rewarding participants. In the BTS, the i-score is usually combined with a prediction-accuracy score. Why is this score not considered here? The reason for choosing one and not the other score is rather implicit at the moment. Also, from the verbal definition on p. 4, the exact calculation is not clear to me: Is this how the score is calculated for each *item*? Or for each *response* on each *item*? Or each *respondent*? Without proper mathematical notation that includes indices denoting respondents, items, and response options, this is unclear. Therefore, I recommend that the authors include a more explicit justification for their choice of measure and replace the verbal equation on p. 4 with a formally correct mathematical equation.

2.2) Relatedly, are participants ranked within each condition or across conditions?

3) Statistical analysis

3.1) My main concern with this Stage 1 report is the specification of the three statistical hypotheses and the proposed analysis plan. In their hypotheses, the authors formulate three pairwise differences that they plan to test these with three two-sided (Welch's) t-tests:

$$H_1: \mu_{RI} < \mu_{BTS},$$

$$H_2: \mu_{RI} < \mu_{BTS+IP},$$

$$\mu_3: \mu_{BTS} < \mu_{BTS+IP}.$$

First of all, I do not understand why the authors would conduct two-sided tests when the hypotheses are clearly directed. Directed hypotheses warrant one-sided tests. Otherwise, the statistical models at tests do not correspond to the substantive hypotheses, resulting in unnecessarily conservative tests.

Moreover, and more importantly, the authors argue that the three tests are based on independent null hypotheses. I do not agree with this assessment, because the hypotheses are clearly not mutually independent: H1 and H3 imply H2. Thus, the tests are in fact redundant and not independent (and neither are the corresponding null hypotheses). The authors formulate precise expectations about the ordering of mean scores across conditions (as noted above, see also #1). I suggest that these expectations be put to the test in a more critical and powerful way, namely, by means of planned contrasts.

In my opinion, there are two substantively relevant contrasts that the authors want to address in their study: (1) Does the BTS in general (i.e., with or without IP) lead to more truthful responding (= higher mean scores) than RI, thus replicating successful prior validations of the procedure? (2) Does the BTS+IP lead to more truthful responding than the BTS without IP? Formally, these correspond to the following orthogonal contrasts Ψ_1 and Ψ_2,

	μ_{RI}	μ_{BTS}	μ_{BTS+IP}
Ψ_1	-2	1	1
Ψ_2	0	-1	1

where Ψ_1 encodes the superiority of BTS versus non-BTS procedures in general, and Ψ_2 further differentiates BTS procedures without versus with IP. In my opinion, a test of these contrasts is a more critical statistical test of the authors' substantive hypotheses, and it is more efficient than three pairwise comparisons.

3.2) *Sample size:* If we assume the same effect size for the two contrasts as the authors assumed for the pairwise comparisons – which makes sense because Ψ_2 is identical to the contrast specified in H3, and the contrast denoted by Ψ_1 should have an even bigger effect when the proposed ordering holds – the required sample size for the same statistical error probabilities ($\alpha = .05, a - \beta = .80$) is much smaller. I include a screenshot of a power analysis in G*Power for this analysis below (where Cohen's f = Cohen's d / 2). Note that in order to calculate the sample size for a one-sided test with $\alpha = .05$, I specified "α err prob = 0.10" in the GUI as it refers to an F-test that is per definition undirected. The resulting sample size to test the above specified, directed contrasts with the desired error probabilities is 620, that is, 207 people per condition (for a two-sided test, which I do not recommend, the required number would be 263 people per condition).

3.3) *Multiverse analysis*: Another suggestion for the analysis plan is to complement the planned frequentist analysis with a Bayesian evaluation by means of Bayes factors. Bayes factors are a continuous measure of statistical evidence for competing statistical hypotheses, which provides informative value beyond a statistical decision, especially in the case of non-significant results. If the authors decide to include Bayesian analysis in the spirit of a "multiverse approach", I recommend for this particular case an approach commonly referred to as "Bayesian informative hypothesis evaluation" (Hoijtink et al., 2019), which is

particularly suitable for testing hypotheses about a specific order of group means. The approach is implemented in the R package bain (https://informative-hypotheses.sites.uu.nl/software/bain/).

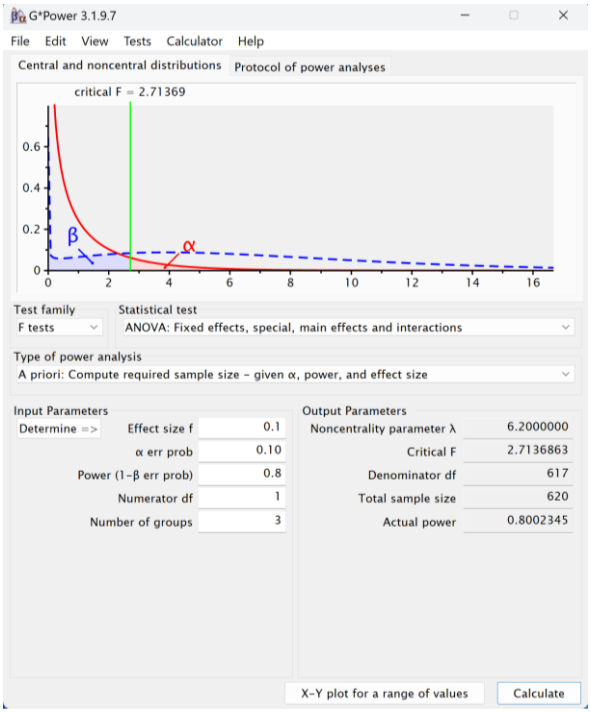

Figure 1: Screenshot of power analysis for planned contrasts in G*Power.

Minor issues:

4) Exclusions/Dropout

The authors adjusted their target sample size to account for potential exclusions. This makes a lot of sense, especially in online studies. However, I wonder whether the estimate of 5% may be too optimistic. It is based on Schoenegger's (2021) findings which, to my knowledge, were not based on 2 study parts and thus, did not include additional dropout. From personal experience, dropout rates in a multi-part studies can be quite high. Thus, the authors may want to increase the proportion of participants to add to the target sample size to account not only for exclusions but also dropout between study parts.

5) Quality check

I found the wording of the quality check potentially misleading. It says, "What percentage of those with the highest information scores will receive a bonus", where 50% is the correct response. However, the phrase "those with the highest information scores" could be perceived as referring to a subset of participants, namely, those that are in the top 50%. In this case, the correct response would be 100%, because all members of this subset should receive the bonus.

Sincerely,
Martin Schnuerch